International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 4, July - August 2016 **ISSN 2278-6856**

A Novel Technique for Finding Influential Nodes

Mini Singh Ahuja

Department of Computer science, Guru Nanak Dev University, Regional Campus, Gurdaspur

Abstract

In complex networks there is a big issue to find the nodes which are playing main role in efficient spreading of information. In this paper, we select a common class of node-level behaviors in which a node is connected to other nodes based on the degree distribution of the node. This choice is encouraged by the network science literature which has shown that many real networks show scale-free structures, which result from preferential attachment. A preferential attachment method belongs to "a class of processes in which some quantity, typically some form of wealth or credit, is distributed among a number of entities or objects according to how much they already have, so that those who are already wealthy get more than those who are not". In the context of network generation models, this "wealth" is the degree of each node.

Keywords: Complex networks, Influential nodes, Degree distribution.

1. INTRODUCTION

Identifying the most influential nodes in complex networks is an important issue for more efficient spread of information or best design of resource allocation [1, 2]. It shows new vision for applications such as finding social leaders [3, 4], influential directors, designing viral marketing strategies, protecting critical regions from intended attacks, and ranking reputation of publications, scientists and athletes. The topology structure performs good in a network's function and behavior, thus it is adequate to rank the nodes according to their function in the network. In fact, not all nodes are equal. Even those with the same degree may function differently in spreading dynamics which makes the problem of finding influential nodes a difficult task. The basic assumption is that given a specific spreading scheme, we rank the nodes according their influence on the range and speed of the spreading. The straightforward way is to greedily assess the size of the outbreak for each node or combination in the network [1, 2]. This approach, however, is so computationally exhaustive that one may not get the result in a reasonable time. This has led to a lot of works focusing on ranking measures, which are supposed to provide objective and quantifiable measures of nodes' importance, from the view of the structural analysis view. It is very common to rank the nodes according to their centrality, which measures how central a node is located in a network. For instance, it is widely believed that the most connected nodes (hubs) are key players in the spreading process and also the nodes with higher betweenness centrality, which measures how many shortest paths cross through this node. The widely used measures of centrality include degree, betweenness, closeness, and eigenvector centrality. In addition, some network-based diffusion algorithms are also used to rank the nodes by taking benefit of the global features of the network [8, 9].

2. MEASURES OF COMPLEX NETWORKS

There are many measures of complex networks. To find the most influential node; some measures of complex networks are discussed here. They are as follow:

- **The average distance**

The average distance is the average, over all the pair of nodes, of the distance between them, i.e. the minimal number of links one has to cross to go from one node to the other. In a random network, it is known that the average distance grows as log(n). The most important distance measure in networks is the minimal number of hops. That is, the distance between two nodes in the network is defined as the number of edges in the shortest path between them. If the edges are supposed to be weighted the lowest total weight path may also be used, which is known as the optimal path [5, 6]. The usual mathematical definition of the diameter of the network is the length of the path between the furthest nodes in the network. The geometrical distance between nodes is worthless. Shortest paths play a significant role in the transport and communication within a network. If one sends a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would attain a fast transfer and save system resources [8]. For such a reason, shortest paths have also played a vital role in the characterization of the internal structure of a graph.

- **The clustering coefficient**

The clustering coefficient is the probability of presence of a connection between two nodes when they are both neighbors of a same node. It is calculated by dividing the total number of triangles (trios of nodes with all the three possible connections) in the network by the total number

of connected triples (trios of nodes with at least two connections). The clustering coefficient is a ratio N / M, where N is the number of edges between the neighbors of n, and M is the maximum number of edges that could possibly exist between the neighbors of n where n is any node in the network . The clustering coefficient of a node is always a number between 0 and 1. The clustering coefficient of a node is the number of triangles that pass through this node, comparative to the maximum number of 3-loops that could pass through the node. Moreover, the clustering coefficient is equal to p since each pair of nodes is linked with the same probability p. This means that, if one considers a group of networks where the average degree is a constant (which is reasonable in the real-world cases), then the clustering coefficient tends to 0 when n grows [7]. The **average clustering coefficient distribution** gives the average of the clustering coefficients for all nodes n with k neighbors for $k = 2$The **network clustering coefficient** that is the average of the clustering coefficients for all nodes in the network.

- **The degree distribution**

The degree distribution is the function P_k giving the proportion of nodes with degree exactly k, i.e. with exactly k neighbors, in the network. In other words, P_k is the probability that a randomly selected node has degree k [1, 2].

3. LITERATURE SURVEY

B.Hou, et al. [1] presented that identifying the most influential nodes in complex networks provides a solid basis for understanding spreading dynamics and ensuring more efficient spread of information. Due to the heterogeneous degree distribution, they noticed that current centrality measures are associated in their results of nodes ranking. They introduced the concept of all-around nodes, which act like all-around players with good performance in collective metrics. Then, an all-around distance is presented for computing the influence of nodes. The experimental results of susceptible-infectious-recovered (SIR) dynamics suggest that the proposed all-around distance can act as a more precise, stable indicator of influential nodes.

D. Chen, et al. [2] presented that identifying influential nodes that lead to faster and broader spreading in complex networks is of theoretical and practical significance. The degree centrality method is very simple but of slight relevance. Global metrics such as betweenness centrality and closeness centrality can better identify influential nodes, but are unable to be applied in large-scale networks due to the computational complexity. In order to design an effective ranking method, they proposed a semi-local centrality measure as a trade-off between the low-relevant degree centrality and other time-consuming measures. They used the Susceptible– Infected–Recovered (SIR) model to evaluate the performance by using the spreading rate and the number of infected nodes. Simulations on four real networks show that they identify influential nodes very well.

J.Zhou e al. [3] states that many significant applications can be generalized as the influence maximization problem, which targets finding a K-node set in a social network that has the maximum influence. Previous effort only considers that influence is propagated through the network with a uniform probability. However, because users actually have different preferences on topics, such a uniform propagation can result inaccurate results. To solve this problem, they have designed a two-stage mining algorithm (GAUP) to mine the most influential nodes in a network on a given topic. Given a set of users' documents considered with topics, GAUP first computes user preferences with a latent feature model based on SVD or a model based on vector space. Then to find top-K nodes in the second stage, GAUP adopts a greedy algorithm that is certain to find a solution within 63% of the optimal. Their evaluation on the task of expert finding shows that GAUP performs better than the state-of-the-art greedy algorithm, SVD-based collaborative filtering, and HITS.

T. Zhu et al. [4] presents that information flows in a network where individuals influence each other. They studied the influence maximization problem of finding a small subset of nodes in a social network that could maximize the spread of influence and proposed a novel information diffusion model CTMC-ICM, which presents the theory of Continuous-Time Markov Chain (CTMC) into the Independent Cascade Model (ICM). Furthermore, they proposed a new ranking metric named Spread Rank generalized by the new information propagation model CTMC-ICM. They experimentally demonstrated the new ranking process that can, in general, extract nontrivial nodes as an influential node set that maximizes the spread of information in a social network and is more efficient than a distance-based centrality.

J. Bae et al. [5] states that identifying influential spreaders is a significant issue in understanding the dynamics of information diffusion in complex networks. The k-shell index, which is the topological location of a node in a network, is a more efficient measure at capturing the spreading ability of a node than are the degree and betweenness centralities. However, the k-shell decomposition fails to produce the monotonic ranking of spreaders because it assigns too many nodes with the same k-shell index. They proposed a novel measure, coreness centrality, for evaluation of the spreading influence of a node in a network using the k- shell indices of its neighbors. Their experiments; on both real and artificial networks, compared with an epidemic spreading model, show that the proposed technique can quantify the node influence more accurately.

Q. Li et al [6] presents that identifying influential spreaders is crucial for understanding and controlling

spreading methods on social networks. Via assigning degree-dependent weights onto links associated with the ground node, they proposed a variant to a recent ranking algorithm named Leader Rank. According to the simulations on the standard SIR model, the weighted Leader Rank executes better than Leader Rank in three aspects the ability to find out more influential spreaders and the higher tolerance to noisy data and the higher robustness to intended attacks.

X. Zhang et al. [7] presents a technique to find a small subset of influential individuals in a complex network such that they can spread information to the largest number of nodes in the network. Though some heuristic methods, including degree centrality, betweenness centrality, closeness centrality, the k-shell decomposition method and a greedy algorithm, can help recognizing influential nodes, they have limitations for networks with community structure. This paper reveals a new measure for assessing the influence effect based on influence scope maximization, which can complement the traditional measure of the expected number of influenced nodes. A novel method for finding influential nodes in complex networks with community structure is proposed. This technique uses the information transfer probability between any pair of nodes and the k-medoid clustering algorithm. The experimental results show that the influential nodes found by the k-medoid method can influence a larger scope in networks with obvious community structure than the greedy algorithm without reducing the expected number of influenced nodes.

C. A.Ruggiero et al. [8] present specific choices about how to represent complex networks. They can have a substantial impact on the execution time required for the respective construction and examination of those structures. In this work they report a comparison of the effects of representing complex networks statically by adjacency matrices or dynamically by adjacency lists. Three theoretical models of complex networks are considered: Erdos–Renyi as well as the Barabasi–Albert model. They examined the effect of the different representations with respect to the construction and measurement of several topological properties (i.e. degree, clustering coefficient, shortest path length, and betweenness centrality). They found that different forms of demonstration generally have a substantial effect on the execution time, with the sparse representation frequently resulting in strangely superior performance.

Z. Sha et al. [9] presented that network design and optimization research has traditionally been concentrated on networks where the designers have direct control over the nodes and their connectivity. However there is increasing importance of social, economic and technical networks whose structures are not under the direct control of the designers, but change as a result of decisions and behaviors of individual self- directed entities. These networks are endogenous in nature, where the local features and behaviors of nodes affect the overall structures. The structure of a network affects its properties, and the properties affect the system's performance. Hence, the problem of designing such endogenously evolving networks involves determining the node-level characteristics and behaviors through appropriate incentives to attain the desired system-level performance. Their aim is to illustrate the problem of designing endogenously evolving networks. They performed a conceptual exploration of the problem, presented the current state of the art and identified the research gaps. The illustrative example involves designing an endogenous network with two objectives, robustness to random node failure and resilience to targeted attack, considering specific node-level characteristics, additional attractiveness, as the design variables. The effect of the design variables on the performance of the network, and potential applications are also discussed.

4. PROPOSED METHODOLOGY

The system performance considered in this paper is the network's robustness against random failure of nodes. For various infrastructure networks such as the Internet, power grids, and transportation networks the robustness of networks is important. On the topology of networks such as Internet the effect of random failure of nodes can be calculated. The nodes are randomly removed from the network and the corresponding effect on the network structure is observed.

In this paper, we choose a general class of node-level behaviors in which a node links to other nodes based on the degree distribution of the node. This choice is inspired by the network science literature which has revealed that many real networks exhibit scale-free structures, which result from preferential attachment. A preferential attachment process belongs to "a class of processes in which some quantity, typically some form of wealth, is distributed among a number of individuals or items according to how much they already have, so that those who are already wealthy receive more than those who are not". In the context of network generation models, this "wealth" is the degree of each node.

Therefore in this paper we have calculated the influential nodes based on the degree of distribution. The nodes with less degree distribution are considered as influential nodes.

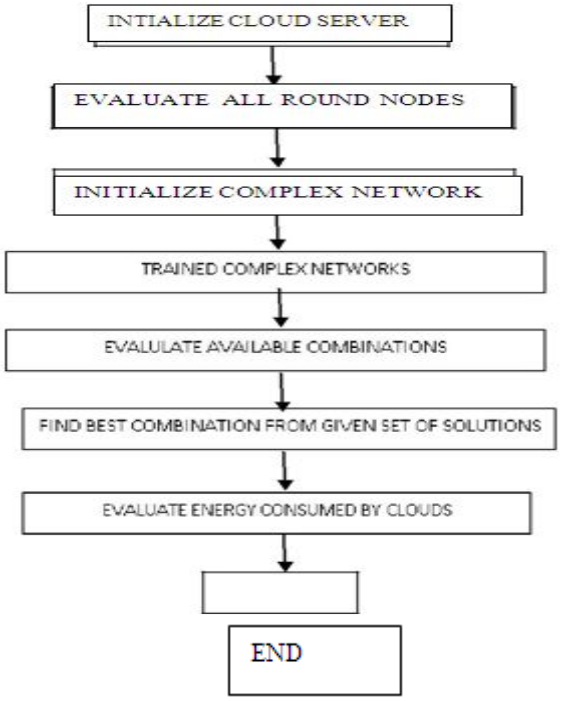
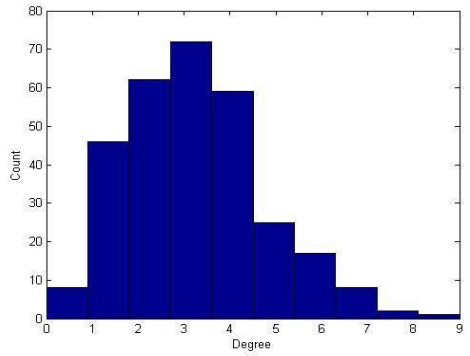
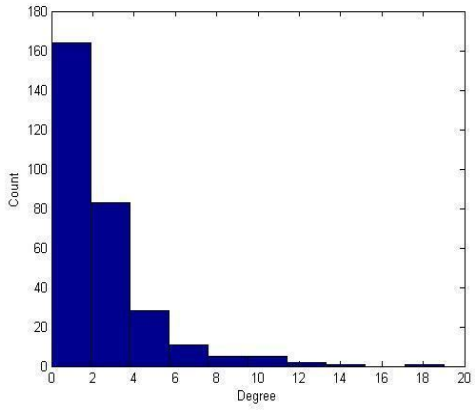


Fig 1: Flowchart of the proposed methodology

5. RESULTS



(a)



(b) existing

Fig2: Degree Distribution (a) proposed technique (b) existing technique

Fig2 shows the degree distribution of nodes in the complex network where degree represent the number of nodes connected to a single node forming the network and count represent the number of nodes of that degree. In the proposed technique fig2(a) , nodes with less degree of distribution are taken as influential nodes also more influential nodes are identified. Figure 2 (b) shows that the existing measure is not giving efficient results for finding influential nodes in complex network.

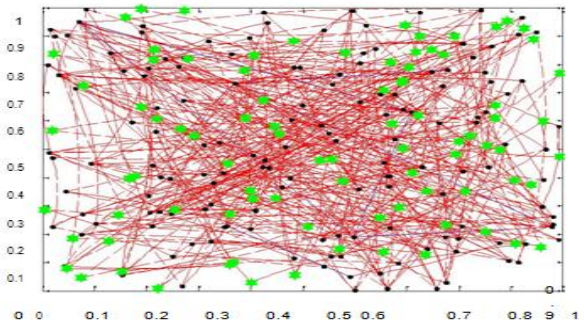


Figure 3: Influential nodes in proposed technique (87)

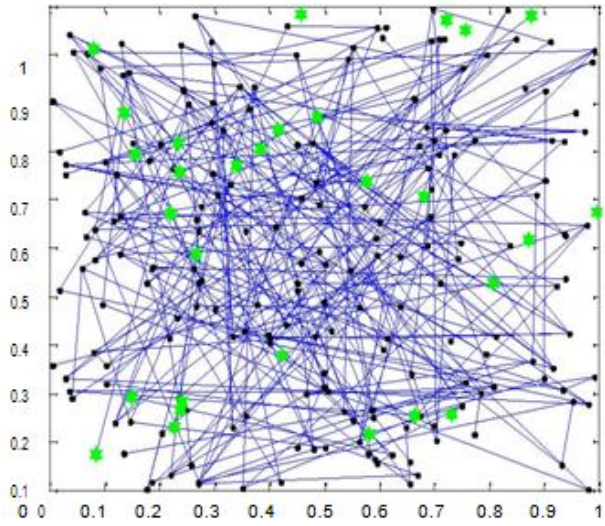


Figure 4: Influential nodes in existing technique (29)

Figure 3 represents the influential nodes in complex network with green color. Total 87 influential nodes are identified by the proposed technique. Figure 4 shows the influential nodes in green which are detected by the existing technique which are very less than the influential nodes found in new ranking measure.

6.CONCLUSION AND FUTURE SCOPE

In this paper, a new ranking measure from a structural view for identifying influential nodes in cloud computing using complex networks has been proposed. The proposed all-around distance provides an optimization for ranking nodes through synthetically combining the degree, betweenness and k-shell ranking measures. When applied to the real-world networks, it effectively finds the more influential spreaders than other ranking measures. This

work also suggests that the all-around distance using complex networks instead of heuristic could be a more effective and stable indicator. The overall objective of this work is to design and implement complex networking based cloud computing to characterize the nodes of cloud computing which are more influential than others.

References

[1]. B.Hou, Y. Yao, and D. Liao ,"Identifying all-around nodes for spreading dynamics in complex networks" In Physica A,,pp.4012-4017,ELSEVIER,2012.

[2]. D. Chena, L. Lu and M.S Shang," Identifying influential nodes in complex networks" In Physica A, pp .1777-1787, ELSEVIER, 2011.

[3]. J.Zhou, Y.Zhang, J.Cheng, "Preference-based mining of top-K influential nodes in social networks"In Future Generation Computer Systems, pp.40-47, ELESVIER, 2014

[4]. T. Zhu, B. Wang , B. Wu and C. Zhu" Maximizing the spread of influence ranking in social networks"In Information Sciences, pp. 535-544, ELESVIER, 2014.

[5]. J. Bae, S. Kim "Identifying and ranking influential spreaders in complex networks by neighbourhood coreness." In Physica A, pp549-559 ELESVIER , 2014

[6]. Q. Li,, T. Zhou, L. Lu and D. Chen "Identifying influential spreaders by weighted Leader Rank" ,In Physica A , pp. 47-55 ,ELESVIER ,2014.

[7]. X. Zhang, J. Zhu, Q. Wang and H. Zhao "Identifying influential nodes in complex networks with community structure. "In Knowledge-Based Systems, pp. 74-84, ELESVIER, 2013.

[8]. C. A.Ruggiero, O. M. Bruno, G. Travieso and L. F. Costa "On the efficiency of data representation on the modeling and characterization of complex networks" In Physica A, pp. 2172-2180, ELESVIER, 2011.

[9]. Z. Sha and J. H. Panchal "Towards the design of complex evolving networks with high robustness and resilience." In

Conference on Systems Engineering Research (CSER'13), pp. 522-531, ELESVIER, 2013.

[10]. Wang, Yujie, L. Xing, and H. Wang. "Reliability of scale-free complex networks." In Reliability and Maintainability Symposium (RAMS), 2013 Proceedings-Annual, pp. 1-6. IEEE, 2013.

[11]. M. Katyal , A. Mishra, A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment ,International Journal of Distributed and Cloud Computing, Volume 1 Issue 2,(December 2013).

[12]. Li, Yun, G Liu, and Song-yang Lao. "Overlapping community detection in complex networks based on the boundary information of disjoint community." In Control and Decision Conference (CCDC), 2013 25th Chinese, pp. 125-130. IEEE, 2013.

[13]. Youssef, Bassant E., and M. RM Rizk. "SNAM: A heterogeneous complex networks generation model." In Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on, pp. 44-50. IEEE, 2014.

[14]. Zelinka, Ivan, D. Davendra, J. Lampinen, R. Senkerik, and M. Pluhacek. "Evolutionary algorithms dynamics and its hidden complex network structures." In Evolutionary Computation (CEC), 2014 IEEE Congress on, pp. 3246-3251. IEEE, 2014.

[15]. Curia, Vincenzo, M. Tropea, P. Fazio, and S. Marano. "Complex networks: Study and performance evaluation with hybrid model for Wireless Sensor Networks." In Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian Conference on, pp. 1-5. IEEE, 2014.